Title	Treatment efficacy of a behaviorally based voice therapy program for female dysphonic patients
Other Contributor(s)	University of Hong Kong.
Author(s)	Lao, Oliveros Florencia
Citation	
Issued Date	2003
URL	http://hdl.handle.net/10722/48808
Rights	Creative Commons: Attribution 3.0 Hong Kong License

Treatment Efficacy of a Behaviorally Based Voice Therapy Program for Female Dysphonic Patients

Florencia Lao Oliveros

A dissertation submitted in partial fulfillment of the requirements for the Bachelor of Science (Speech and Hearing Sciences), The University of Hong Kong, May 2, 2003.

ABSTRACT

The study investigated the effectiveness of a behaviorally based voice therapy program for Cantonese female dysphonic adults using controlled group study design. Forty-two subjects were assigned to a treatment group and 42 subjects received no treatment. Instrumental and subjective measures were carried out before and after treatment to evaluate changes in vocal quality over time. Results revealed a significant difference between the two groups in the amount of change for the voice severity, the Voice Activity and Participation Profile, fundamental frequency and some acoustic measures ($p<0.01$). Aerodynamic and some acoustic measures failed to show significant differences between the two groups. Discriminant function analysis was used to determine the accuracy in predicting changes in roughness and breathiness after treatment using a minimum number of predicting variables. Results showed the selected predictor variables using stepwise entry method and Pearson product-moment correlations were more accurate in predicting changes in breathiness than in roughness.

INTRODUCTION

A number of voice therapy techniques for treating different types of voice disorders have been studied for their effectiveness. Although the different techniques were found to be able to optimize vocal behavior in reducing trauma to the vocal folds as measured by various instrumental and perceptual measures approaching normal after treatment, many of these studies have methodological flaws which reduced their robustness (Pannbacker, 1998). For example, very few studies adopted controlled group study design in evaluating treatment outcome.

Verdolini-Marston, Burke, Lessac, Glaze, and Caldwell (1995) investigated the effects of a two-week intensive program of confidential voice therapy (N=5) and resonant voice therapy (N=3) for treating laryngeal nodules while comparing with a control group with no treatment (N=5). Pre- and post-therapy phonatory effort were rated on a magnitude estimation scale ranging from 50 (very easy phonation) to 200 (very effortful phonation). Auditory-perceptual ratings ranged from 1 (healthy voice) to 5 (severely impaired voice) and visual-perceptual measures ranged from 1 (healthy vocal folds) to 5 (severe nodules/polyps or related). Significant improvement was shown for the two therapy groups but not for the control group. The authors concluded there was evidence of benefit from both types of therapies as compared to control conditions.

Carding and Horsley (1992) studied the effects of an eight-week treatment program of direct voice therapy (N=10) and indirect voice therapy (N=10) for treating non-organic dysphonia while comparing with a no-treatment group (N=10). Pre- and post-therapy perceptual voice evaluation was based on the use of Buffalo II Voice Profile (Wilson, 1987). Visual-perceptual measures were performed by a laryngologist in the form of medical notes. Electrolaryngography judgements were based on expected visual norms as described by Reed (1982; cited in Carding & Horsley, 1992). Mean speaking fundamental frequency results were compared to those of the normal subjects. A patient questionnaire was also included for the patients to consider their own vocal performance. Significant improvement was shown for the direct voice therapy groups but not for the control group in all the measures. Half of the patients in the indirect voice therapy group showed improvement after treatment. The authors concluded that there was evidence of benefit from direct voice therapies as compared to control conditions.

In another study, Carding, Horsley, and Docherty (1999) investigated the effects of indirect therapy (N=15) and direct therapy with indirect therapy (N=15) on non-organic dysphonia against a control no-treatment group (N=15). Pre- and post-therapy measures used were similar to those of Carding and Horsley (1992) except with the use of a seven-point perceptual rating scale of 1 (normal) to 7 (very severe). Results revealed significant improvement for the direct voice therapy group but not the control group.

Roy, Gray, Simon, Dove, Corbin-Lewis, and Stemple (2001), evaluated the functional effects of vocal hygiene (N=20) against vocal function exercise (N=19) and no-treatment control conditions (N=19) on 58 dysphonic teachers for a six-week period based on subjective evaluations. Treatment efficacy was evaluated based on the Voice Handicap Index (VHI) (Jacobson, Johnson, Grywalski, Silbergleit, Jacobson and Benninger, 1997) and a patient questionnaire about effectiveness of treatment. Only the group receiving vocal function exercise treatment experienced a significant reduction in VHI scores and reported more improvement in overall voice quality. The authors concluded that direct vocal function exercise was more effective than vocal hygiene in reducing dysphonia in the teaching population. However, no objective measures were used in this study.

The four studies reviewed above were the only recent studies that used controlled group study design in studying treatment efficacy of behaviorally based voice therapy program on adult patients. However, they included small subject size or adopted only a single measurement to document treatment efficacy which affected the reliability of the study. Since the exclusive use of only one outcome measure might produce misleading results, a multidimensional approach in documenting outcome was used in the present study. Each measurement technique was evaluated separately below.

Outcome Measures

Acoustic Analyses

Investigators once believed that the mostly widely used acoustic measures e.g. noise-to-harmonic ratio (NHR) and perturbation (e.g. jitter, shimmer) might replace the subjective perceptual voice evaluation. However, recent studies showed that these three measures alone might not be successful in discriminating among dysphonic voice qualities (Bielamowicz, Kreiman, Gerratt, Daucer, & Berke, 1996; Hillenbrand, Cleveland, & Erickson, 1994; Titze, 1995; Yiu, 1999). Hillenbrand et al. (1994) and Hillenbrand and Houde (1996) proposed five periodicity measures which they showed to be good predictors of breathiness. They are:

1. Cepstral Peak Prominence (CPP), which is a measure of cepstral peak amplitude normalized for overall amplitude. The rationale behind this measure is that a highly periodic signal should show a well-defined harmonic structure and, consequently, a more prominent cepstral peak than less periodic signal (Hillenbrand et al., 1994).
2. CPP-smoothed (CPPS), which is a modified measure of CPP by smoothing the individual cepstra before extracting the cepstral peak and calculating the peak prominence.
3. Pearson *r* at autocorrelation peak (RPK), which is a measure of the degree of periodicity, calculated between the signal and itself at a delay. The rationale behind is that a highly periodic signal should show more prominent autocorrelation peaks than less periodic signals. A perfectly periodic signal would show a correlation of 1.0, and the correlation decreases systematically with less periodic signals (Hillenbrand et al., 1994).
4. Breathiness Index (BRI), which is a measure of the ratio of the relative energy of the high-frequency component to the total pre-emphasized voice signal. It is based on the speculation that the degree of breathiness is determined by the relative intensity of the turbulent noise in the higher portion of the frequency range to the total voice signal. The rationale is that breathy voice would show more high frequency energy and thus larger BRI values (Hillenbrand et al., 1994).
5. First and Second Harmonic Amplitude (H1H2), which is a simple measure of the dB amplitude of the first harmonic relative to the second harmonic of a voice signal.

In addition to the periodicity measures, Callan, Kent, Roy, and Tasko (1999) and Fex, Fex, Shiromoto, and Hirano (1994) found that two acoustic measures were able to distinguish dysphonic voices (functional and spasmodic dysphonia) from normal voices. They include:

1. Coefficient of variation of the peak-to-peak amplitude (VAM), which is a measure which looks at the consecutive cycle to cycle differences in the amplitude of a voice signal waveform.
2. Standard deviation of the fundamental frequency (STD), which is the standard deviation of all extracted period-to-period fundamental frequency values (in hertz) within a one-second voice sample.

In addtiton to the seven acoustic measures mentioned above, NHR, jitter and shimmer percent were also used in the present study because they were widely used for clinical and research situations. Connected speech voice samples were used as they were more representative of daily voice use situation. It has been shown that the Computerized Speech Lab (CSL)'s Multi-Dimensional Voice Program (MDVP Model 4300B, Kay Elemetrics Corp.) was able to accurately analyze the variation of acoustic properties in connected speech (Yiu, Worrall, Longland, & Mitchell, 2000).

Aerodynamic Measurements

Aerodynamic measures using mean airflow rate and subglottal pressure were shown to be statistically accurate in discriminating dysphonic and normal voices but it was found that there were high variabilities between and within subjects among all aerodynamic measures (Goozee, Murdoch, Theodoros, & Thompson, 1998; Higgins, Netsell, & Schulte, 1994). To reduce these variabilities, the present study used the average of five speech samples for each measure because an increased number of trials for each measurement should provide a more reliable and representative average value, thus more accurate aerodynamic measurement (Higgins et al., 1994).

Voice Range Profile (Phonetogram)

As phonetogram defines vocal limits by plotting fundamental frequency (F_0) by intensity and evaluating an individual's entire vocal range, it would show treatment effectiveness when vocal limitation reduced for an individual following voice therapy. The present study used the F_0, sound pressure level (SPL) and speaking profile area as measurements in evaluating treatment effectiveness.

Functional Approach in Measuring Treatment Outcome

A multidimensional approach in studying treatment outcome should include patients' perceptions of their voice problems. The Voice Activity and Participation Profile (Ma & Yiu, 2001), a self-assessing questionnaire to evaluate the impact of voice problems on quality of life and functional status in communication of voice patients, has been shown to be effective in distinguishing dysphonic from non-dysphonic individuals. Ma and Yiu (2001) found no correlation between the VAPP and the measure of voice impairment from acoustic analysis and auditory-perceptual ratings as the severity of impairment did not predict the degree of activity limitation or participation restriction. They therefore argued that clinical management should incorporate patient's own perspective of impact on daily function in addition to clinician's point of view towards impairment.

Auditory-Perceptual Ratings

Kreiman, Gerratt, Kempster, Erman, and Berke, (1993) reported that inter- and intra-rater agreement in perceptual voice evaluation ranged from 55% to 100% in the literature. To increase intra- and inter-rater agreement, Kreiman et al. suggested the use of external voice anchors and more sensitive scales for the evaluations. In the present study, a training-program using external synthesized anchors were used for the evaluation procedure.

Although many outcome studies of voice therapy used a range of instrumental measures and subjective evaluations to measure treatment efficacy with small subject groups, the majority of them did not include a control no-treatment subject group to evaluate treatment outcome. So, the present study used a large group randomized control study design (N=84) with multiple outcome measures to evaluate the effectiveness of a behaviorally based voice treatment program.

Research Questions and Hypothesis

The present study had two objectives. The first objective set out to evaluate the effectiveness of a voice therapy program for female dysphonic patients with varying levels of severity of voice disorders. The voice therapy program included vocal hygiene, relaxation exercises and humming techniques (see Yiu & Ma, 2001) because these methods were used most extensively by speech therapists in Hong Kong. It was hypothesized that the different vocal facilitation techniques should reduce trauma to the vocal folds by adopting proper phonation techniques and significant improvement was expected following treatment. Long-term effectiveness of the treatment group was also expected in which no significant changes should be found in the different subjective and objective measurements during the maintenance phase as it was expected that subjects in the treatment group should be able to adopt a proper phonatory mechanism learnt from treatment after therapy.

The second objective of this study was to determine whether a minimal set of voice measures taken before the treatment could accurately predict changes in voice quality after treatment. Discriminant function analysis was used to achieve this objective.

METHOD

Participants

Eighty-four Chinese women with varying types of vocal pathologies served as subjects. They were attending the Voice Clinic of The University of Hong Kong. Forty-two subjects were assigned to a treatment group and 42 were assigned to a no-treatment control group. Subjects in the no-treatment group were on a waiting list for voice therapy (see Appendix I for subjects' occupational background).

Assessment Schedule

Each subject in the treatment group was assessed before treatment (Pre), one week after treatment (Post-1) and four weeks later (Post-2) to evaluate long-term maintenance. For the control group, assessment one (Ax-1) was taken at the same time with the treatment group and assessment two (Ax-2) was taken two months after. Each assessment included administration of VAPP, recordings of speech samples for acoustic, aerodynamic, voice range profile and auditory-perceptual analyses. Details will be described below.

Procedures

Subjects from the treatment group received voice therapy according to the procedures described in the Voice Therapy Instructional Manual (Yiu & Ma, 2001). The voice therapy protocol consisted of three behaviorally based approaches: vocal hygiene, relaxation exercises and humming techniques. The therapy lasted for ten weeks with one 45-minute long therapy session per week. The voice therapy took place at the Voice Clinic of The University of Hong Kong by a qualified speech and language pathologist of the Voice Research Laboratory. Yiu and Ma (2001) provide detailed session-to-session information of the treatment program.

Data Collection and Analysis

Acoustic Analysis

For acoustic analysis, the jitter %, shimmer %, noise-to-harmonic-ratio (NHR), coefficient of variation of the peak-to-peak amplitude (VAM), and standard deviation of the fundamental frequency (STD) were obtained from the average of five speech sample of the Cantonese sentence /pa pa ta pɔ/ (father hits ball). The analyses were performed using Computer Speech Lab's Multidimensional Voice Profile (MDVP Model 4300B, Kay

Elemetrics Corp.). The recordings for CPP, CPPS, RPK, H1H2 and BRI measures were obtained from the average of five sustained /a/ phonations. The analyses were performed using Hillenbrand's Data Analysis (Hillenbrand et al., 1994; Hillenbrand & Houde, 1996). CPP, CPPS and BRI were fully automated with the software developed by Hillenbrand et al.. CPP, CPPS and RPK were calculated from signals that were bandpass filtered between 2.5 and 3.5 KHz, and highpass filtered at 2.5 KHz. BRI was derived from unfiltered signals.

Aerodynamic Analyses

Aerodynamic analyses were performed using Kay Elemetrics' Aerophone II (Model 6800, Kay Elemetrics Corp.). Measures obtained included mean flow rate of 1) maximum sustained /a/ production, 2) maximum sustained /i/ production, 3) maximum sustained /u/ production, 4) vowel /a/ phonation for a comfortable period of time, 5) sentence production /pa pa ta pɔ/, 6) peak air pressure from sentence production /pa pa ta pɔ/, and 7) peak air pressure from a vocal efficiency task by producing strings of vowel-consonant syllables /ipipipipipi/.

The mean flow rate for each vowel /a/, /i/ and /u/ in the maximum sustained phonation tasks and vowel /a/ in the most comfortable phonation tasks were calculated by including the lowest point of the rising slope and the lowest point of the falling slope on the sound pressure waveform display. The peak intra-oral pressure measurement was based on the middle five /pi/s in the vowel-consonant string production. This was carried out by identifying the lowest point of the falling slope of the fuse sound pressure peak and the lowest point of the falling slope of the sixth sound pressure peak on the waveform display. The airflow and intra-oral pressure measurements from the sentence production task were done by extracting the lowest point of the first sound pressure rising slope and the lowest point of the falling slope of the fourth peak on the waveform display.

Voice Range Profile (Phonetogram)

The comfortable speaking voice of each subject reading a Cantonese passage "North Wind and the Sun" were captured by the Swell's real-time computerized phonetogram Phog 1.0 from AB Nyvalla DSP. Measures obtained included 1) maximum fundamental frequency 2) minimum fundamental frequency 3) speaking fundamental frequency range 4) maximum speaking intensity 5) minimum speaking intensity 6) speaking intensity range and 7) speaking profile area. Boundary values such as maximum-minimum frequency and intensity values were analysed by clicking the mouse cursor on the corresponding locations of the voice profiles. All these values were calculated by the software automatically. Frequency range (in semitone) and intensity range (in dBA) of voice signal were calculated from the maximum-minimum frequency and intensity values.

Functional Measures

Each participant was administered the VAPP (Ma & Yiu, 2001). Scores such as total Activity and Limitation Score (ALS) and Participation Restriction Score (PRS) for the Job, Daily Communication and Social Communication section, Emotion score, total ALS, total PRS and total VAPP score were obtained.

Auditory-perceptual Evaluation

Recordings for auditory-perceptual analyses and listening material. The recording signals were digitized and stored using Kay Electronics' Computer Speech Lab. A sentence production task was chosen for perceptual analyses because connected speech resembled daily speaking situation most. For the treatment group, sentences produced during Pre, Post-1 and Post-2 served as the sources for listening materials. For the no-treatment group, speech samples from Ax-1 and Ax-2 served as listening material.

Raters. Three female listeners were recruited from among the final Year Four students in the Division of Speech and Hearing Sciences, Faculty of Education at The University of Hong Kong. All listeners were native speakers of Cantonese with normal hearing abilities and mean age of 22.

Auditory-perceptual training program. To minimize listener variation, all the raters were required to undergo a 30-minute perceptual voice training program prior to carrying out the perceptual evaluations. The training program used synthesized voice signals for rating breathiness and roughness. The rating procedures were identical to the actual perceptual evaluation task which will be presented in detail in the following section. The training focused on breathiness and roughness only and no synthesized signals were available for the overall severity.

The anchors used were based on a prototype sentence developed by Yiu, Murdoch, Hird, and Lau (2002). Voice signals for roughness were based on signals created using synthesizer HLSyn version 2.2 (Sensimetrics) by varying the parameter amplitude of voicing, amplitude of aspiration and diplophonia. Breathiness signals were based on signals created by varying the parameter amplitude of aspiration and diplophonia. In the training program, after the listener had given a rating for each quality, an answer was provided for the corresponding quality. A separate program was available each for training roughness and breathiness.

Auditory-perceptual evaluation procedures. Each listener participated in a two-hour rating session immediately after the 30-minute training session. Listeners were informed that the speakers of the stimuli were female Cantonese speakers who had been recorded several times. No other information was given about the background and the purposes of the study.

Kreiman, Gerratt, Precoda, and Berke (1992) adopted paired comparison in judging dissimilarities between two voice signals by asking listeners to rate the dissimilarity of voices on a seven-point equal-appearing interval scale. The present study used a similar paired comparison method where the raters were asked to judge the dissimilarities of two voice signals at a time. Voice signal pairs were presented using a computer with Microsoft PowerPoint over a pair of headphones (model HD25, Sennheiser) through a Creative Sound Blaster Extigy System sound card in a sound-treated booth. Each listener evaluated the voice pairs individually on a rating form. They were asked to listen carefully to each pair and to rate the dissimilarity of the voices on an equal-appearing interval scale ranging from –7 to 7. For every voice pair, they were asked to rate the second signal (signal B) while using the first signal (signal A) as a reference. For example, if signal B was much more breathy than signal A, they were asked to rate signal B as 7. If signal B had the same degree of breathiness as signal A, they were asked to rate signal B as 0. If signal B was much less breathy than signal A, they were asked to rate signal B as –7. Three parameters, roughness, breathiness, and overall severity were marked on separate scales. The order of the first and second stimuli were semi-randomized, for example, 50% of the posttreatment stimuli were presented first in a pair and 50% of the pretreatment stimuli were presented first in a pair. Order of presentation of the voice sets was randomized across listeners.

Reliability. Agreement data was used to evaluate inter-rater and intra-rater reliability. For intra-rater reliability, 20% of the total voice signals were repeated and incorporated into the original set of stimuli to be rated.

Statistical Analysis

Since a large number of statistical tests were adopted, a more stringent alpha level (0.01) was used to avoid inflating Type I and Type II errors.

Independent Variables for Predicting Vocal Quality Changes Following Voice Therapy

Since perceptual voice quality ratings often serve as standards against objective measures of voice qualities (Holmberg, Hillman, Hammarberg, Sodersten, & Doyle, 2001), changes in roughness and breathiness after therapy were chosen as dependent variables in the discriminant function analysis. Although changes in overall severity should describe the entire vocal quality, it was not chosen as dependent variable because the training program offered to the listeners did not include this quality. It was assumed that roughness and breathiness should be more reliable measures.

The predictor variables were chosen from all instrumental and subjective measures which have shown significant changes across time in the treatment group. Two discriminant function analysis were carried out, one for roughness and one for breathiness. The selecting procedure was based on either one of the following criteria: 1) the selected parameter should either have pretreatment scores that correlated significantly with the dependant variable or 2) it should be selected as having discriminating power in predicting vocal quality changes after treatment using the stepwise entry method of discriminant function analysis. The appropriate measures were selected separately for breathiness and roughness. Pearson product-moment correlation coefficient was used to evaluate the correlations of the 14 outcome measures that had shown significant changes across time for the treatment group and the dependent variables (changes in roughness and breathiness after treatment). These 14 outcome measures were based on results obtained from all the instrumental and subjective outcome measures that will be discussed in the next session. Results from the Pearson product-moment correlation coefficient test and the stepwise entry method of discriminant function analysis are presented in Appendix II and III respectively.

For roughness, the following measure showed statistically significant correlation with changes in roughness after treatment (see Appendix II):

1) Maximum fundamental frequency

Based on the stepwise entry method, the following measure was selected as having discriminating power in predicting changes in roughness after treatment (see Appendix III):

1) Pearson *r* at autocorrelation peak

To include a variety of parameters in the predicting variable cluster, Total VAPP was selected because it represented patient's perspective towards dysphonia, though its correlation with roughness was not significant (see Appendix II).

For breathiness, the following measure showed statistically significant correlation with changes in breathiness after treatment (see Appendix II):

1) Pearson r at autocorrelation peak

Based on the stepwise entry method, the following measures were selected as having discriminating power in predicting changes in breathiness after treatment (see Appendix III):

1) Mean flow rate for sentence production

2) NHR

3) Self perceived severity

Although Self perceived severity was selected as having discriminating power in predicting changes in breathiness, to be consistent with the predictor variables selected for roughness, total VAPP was used to replace Self perceived severity in the discriminant function analysis to determine the accuracy in predicting vocal quality changes after treatment.

RESULTS

Aerodynamic Measures

The number of subjects in each group varied for each measurement because the subjects produced some unanalyzable aperiodic signals.

Treatment Effect

Tables 1 and 2 list the descriptive statistics for the treatment and no-treatment groups respectively. No significant changes were found for the measurements in mean flow rates and peak intraoral pressures for the treatment and no-treatment groups across time (p>0.04, see Appendix IV A). To determine whether the two groups demonstrated significant changes over time, the difference between Pre and Post-1 measurements for the treatment group and the difference between Ax-1 and Ax-2 measurements for the no-treatment group were analyzed. No significant difference was found between the two groups for all seven aerodynamic measures (p>0.1, see Appendix IV B).

Table 1. Descriptive statistics for the aerodynamic measures for the treatment group.

		Pre			Post-1			Post-2		
Aerodynamic measures	**#N**	**Mean**	**SD**	**Range**	**Mean**	**SD**	**Range**	**Mean**	**SD**	**Range**
Mean flow rate (l/s)										
Maximum sustained /a/	42	0.14	0.07	0.04–0.35	0.15	0.12	0.04–0.58	0.13	0.06	0.04–0.28
Maximum sustained /i/	42	0.16	0.11	0.04–0.46	0.15	0.10	0.04–0.52	0.12	0.08	0.03–0.35
Maximum sustained /u/	42	0.19	0.15	0.03–0.68	0.14	0.07	0.02–0.36	0.17	0.12	0.05–0.63
Most comfortable /a/	42	0.21	0.17	0.20–0.76	0.20	0.17	0.03–0.81	0.15	0.09	0.02–0.35
Sentence	30	0.28	0.23	0.01–1.18	0.26	0.17	0.04–0.66	0.14	0.09	0.03–0.38
Peak intraoral (subglottal) pressure (cmH_2O)										
Sentence	16	8.68	2.44	4.48–15.22	9.04	2.20	4.69–13.86	9.20	0.09	2.11–26.62
Vowel-consonant string	25	9.54	3.86	2.08–18.72	11.15	4.07	5.20–21.71	12.25	4.67	2.22–24.68

*Significant at 0.01 level (two-tailed). #Group size varied for each measurement because the subjects produced some unanalyzable aperiodic signals. Pre=pretreatment. Post1=posttreatment one. Post-2=posttreatment two. l/s=litre per second. cmH_2O=centimeters of water. SD=standard deviation.

Table 2. Descriptive statistics for the aerodynamic measures for the no treatment group.

		Assessment-1			Assessment-2		
Aerodynamic measures	**#N**	**Mean**	**SD**	**Range**	**Mean**	**SD**	**Range**
Mean flow rate (l/s)							
Maximum sustained /a/	41	0.17	0.10	0.05–0.39	0.16	0.09	0.03-0.41
Maximum sustained /i/	42	0.16	0.08	0.04–0.39	0.16	0.12	0.04–0.76
Maximum sustained /u/	38	0.19	0.09	0.07–0.38	0.18	0.08	0.06–0.47
Most comfortable /a/	41	0.18	0.11	0.04–0.68	0.16	0.08	0.04–0.49
Sentence	39	0.18	0.13	0.05–0.71	0.43	1.66	0.05–10.65
Peak intraoral (subglottal) pressure (cmH_2O)							
Sentence	22	10.51	4.21	5.13–20.94	10.52	3.56	5.94–19.14
Vowel-consonant string	20	13.65	4.99	5.74–23.56	15.19	5.58	6.50–32.87

#Group size varied for each measurement because the subjects produced some unanalyzable aperiodic signals. l/s=litre per second. cmH_2O=centimeters of water. SD=standard deviation.

Maintenance Effect

A significant improvement was, however, demonstrated in the mean flow rate in sentence production from Post-1 to Post-2 measurements (paired t=3.72, two-tailed, p=0.001) for the treatment group (see Appendix IV A, Post-1 vs. Post-2 column).

Acoustic Analysis

Treatment Effect

Tables 3 and 4 list the descriptive statistics of the acoustic analysis (MDVP) for the treatment and no-treatment groups. Significant improvement was found for jitter and shimmer percent for the treatment group. No significant changes were found for the no-treatment group across time (see Appendix IV C).

Table 3. Descriptive statistics for acoustic analysis (MDVP) of the treatment group.

Acoustic		Pre			Post-1			Post-2		
Measures	#N	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
VAM	30	44.07	11.64	28.22-83.80	38.88	9.88	24.04-60.06	36.93	7.34	28.54-56.05
STD	30	29.50	10.14	15.04-51.73	34.18	16.02	10.48-84.71	28.59	8.13	11.64-49.76
Jitter %	30	3.57	1.23	1.59-6.46	2.89	1.22	1.33-6.69	2.70	1.31	1.12-6.20
Shimmer %	30	9.56	2.89	5.69-15.80	7.33	2.23	4.23-14.35	7.35	2.65	3.53-14.78
NHR	30	0.29	0.10	0.14-0.49	0.24	0.09	0.15–0.59	0.23	0.06	0.14-0.40

#Group size varied for each measurement because the subjects produced some unanalyzable aperiodic signals. Pre=pretreatment. Post-1=posttreatment one. Post-2=posttreatment two. VAM=Coefficient of Variation of Peak to Peak Amplitude. STD=Standard Deviation of the Fundamental Frequency. NHR=Noise to Harmonic Ratio.

Table 4. Descriptive statistics for acoustic analysis of the no-treatment group.

Acoustic		Assessment-1			Assessment-2		
Measures	#N	Mean	SD	Range	Mean	SD	Range
VAM	34	39.60	6.71	26.14 – 52.13	39.86	6.17	25.75 – 51.44
STD	34	35.68	12.80	17.86 – 71.02	31.34	11.66	12.04 – 62.21
Jitter%	34	3.35	1.16	1.64 – 6.86	3.41	1.29	1.70 – 6.24
Shimmer%	34	9.19	2.22	5.96 – 15.72	9.82	2.46	6.68 – 16.77
NHR	34	0.26	0.05	0.17 – 0.42	0.26	0.05	0.17 – 0.42

#Group size varied for each measurement because the subjects produced some unanalyzable aperiodic signals. VAM=Coefficient of Variation of Peak to Peak Amplitude. STD=Standard Deviation of the Fundamental Frequency. NHR=Noise to Harmonic Ratio.

In order to determine whether the two groups of subjects demonstrated significant changes over time, the difference between the Pre and Post-1 measurements for the treatment group and Ax-1 and Ax-2 measurements for the no-treatment group were analyzed. Significant differences were found between the two groups for shimmer percent and NHR

(p<0.01, see Appendix IV D). No significant differences were found for VAM, STD and jitter percent (p>0.013, see Appendix IV D).

Tables 5 and 6 list the descriptive statistics of acoustic measures based on Hillenbrand and Houde (1996) for the treatment and no-treatment groups respectively. No significant changes were found for both groups across time (p>0.02, see Appendix IV E). No significant difference was found between the difference from Pre and Post-1 measurements for the treatment group and the difference from Ax-1 and Ax-2 measurements for the no-treatment group over time (p>0.2, see Appendix IV F).

Table 5. Descriptive statistics for acoustic measures based on Hillenbrand and Houde (1996) for the treatment group.

Acoustic		**Pre**			**Post-1**			**Post-2**		
Measure	**#N**	**Mean**	**SD**	**Range**	**Mean**	**SD**	**Range**	**Mean**	**SD**	**Range**
BRI	33	301.61	250.56	39.90-985.60	243.17	136.20	39.99-678.31	283.15	144.64	21.21-639.69
H1H2	38	10.49	6.90	0.82-29.14	10.41	7.61	2.72- 41.60	10.63	5.65	1.91-26.33
CPP	35	14.61	3.79	9.23-22.76	14.67	2.92	10.26-21.63	14.21	2.52	9.52-20.08
CPPS	34	7.76	3.90	0.15-14.81	7.89	2.56	3.22-12.89	7.14	2.34	1.82-12.47
RPK	33	0.84	0.22	0.05-1.00	0.85	0.13	0.57-1.00	0.78	0.15	0.46-1.00

*Significant at 0.01 level (two-tailed); #Group size varied for each measurement because the subjects produced some unanalyzable aperiodic signals; Pre=pretreatment; Post-1=posttreatment one; Post-2=posttreatment two; BRI=Breathiness Index; H1H2=First and Second Harmonics; CPP=Cepstral Peak Prominence; CPPS=Cepstral Peak Prominence-Smoothed; RPK=Pearson r at Autocorrelation Peak.

Maintenance Effect

For the two parameters where significant improvement was observed in the treatment group: jitter and shimmer percent, there were no significant changes from the Post-1 to Post-2 measurements. This indicated that long-term maintenance in voice quality improvement was maintained five weeks after therapy (see Appendix IV C, Post-1 vs. Post-2 column). For the breathiness measurements (Hillenbrand & Houde, 1996), significant improvement was, however, demonstrated in Pearson r at autocorrelation peak (RPK) for the treatment group from Post-1 to Post-2 measurements (see Appendix IV E, Post-1 vs. Post-2 column).

Table 6. Descriptive statistics for acoustic measures based on Hillenbrand and Houde (1996) for the no-treatment group.

Acoustic Measures	#N	Assessment 1 Mean	SD	Range	Assessment 2 Mean	SD	Range
BRI	40	325.89	131.17	46.52 – 602.25	317.92	147.27	78.01– 27.87
H1H2	36	6.35	3.84	2.24 – 18.58	7.46	3.68	1.48 – 15.18
CPP	40	13.93	2.70	9.97 – 20.00	14.04	2.84	9.67 – 21.95
CPPS	40	6.99	2.43	2.77 – 12.66	7.16	2.51	1.47 – 12.81
RPK	39	0.77	0.10	0.55 – 0.99	0.78	0.10	0.45 – 0.99

*Significant at 0.01 level (two-tailed); # Group size varied for each measurement because the subjects produced some unanalyzable aperiodic signals; BRI=Breathiness Index; H1H2=First and Second Harmonics; CPP=Cepstral Peak Prominence; CPPS=Cepstral Peak Prominence-Smoothed; RPK=Pearson *r* at Autocorrelation Peak.

Voice Range Profile (Phonetogram)

Treatment Effect

Tables 7 and 8 list the descriptive statistics for the Voice Range Profile for the treatment and no-treatment groups respectively. Significant improvement was found for maximum fundamental frequency and fundamental frequency range for the treatment group across time (see Appendix IV G). No significant changes were found for the no-treatment group in the seven frequency and intensity measures (p>0.3, see Appendix IV G).

Significant difference between the treatment and no-treatment groups over time was found in maximum fundamental frequency (p=0.002, see Appendix IV H). No significant difference was demonstrated in the two groups for minimum fundamental frequency, fundamental frequency range, maximum SPL, minimum SPL, SPL range and speaking profile area (p>0.02, see Appendix IV H).

Maintenance Effect

The improvement in maximum fundamental frequency and fundamental frequency range was maintained in the treatment group as evident from a lack of significant changes from the measurements taken during Post-1 and Post-2 (five weeks after treatment) (p=0.536 and p=0.840 respectively, see Appendix IV G, Post-1 vs. Post-2 column).

Table 7. Descriptive statistics for Voice Range Profile (phonetogram) for the treatment group.

	Pre				Post-1			Post-2		
Frequency and intensity measures	**N**	**Mean**	**SD**	**Range**	**Mean**	**SD**	**Range**	**Mea**	**SD**	**Range**
Maximum fundamental frequency (Hz)	42	723.00	266.83	277.00–1568.00	912.80	310.93	311.00–1568.00	875.4	269.55	311.00–568.00
Minimum fundamental frequency (Hz)	42	112.68	29.94	52.00–185.00	107.70	35.45	49.00–185.00	103.6	34.68	49.00–185.00
Fundamental frequency range (semitone)	42	31.85	90.20	13.00–48.00	37.18	8.99	14.00–56.00	37.2	8.01	14.00–52.00
Maximum SPL (dBA)	42	111.42	1.23	83.00–120.00	115.79	5.42	101.10–120.00	117.0	4.29	105.20–120.00
Minimum SPL (dBA)	42	71.52	12.94	40.60–90.90	76.34	8.17	57.60–55.30	76.8	7.08	64.40–92.50
SPL range (dBA)	42	39.90	8.03	24.90–58.40	39.44	8.07	25.20–55.30	40.2	8.36	21.70–55.60
Speaking Profile Area (dBA semitone)	42	447.33	167.45	211.30–863.10	509.45	230.50	45.20–1071.10	535.7	195.34	186.40–1264.20

*Significant at 0.01 level (two-tailed). Pre=pretreatment; Post-1=posttreatment one. Post-2=posttreatment two. SPL=sound pressure level. dBA=sound pressure level measured in a-weighting scale. Hz=hertz

Table 8. Descriptive statistics for Voice Range Profile (phonetogram) for the no treatment group.

	Assessment 1				Assessment 2		
Frequency and intensity measures	**N**	**Mean**	**SD**	**Range**	**Mean**	**SD**	**Range**
Maximum fundamental frequency (Hz)	42	848.32	293.46	277.0 - 1319.00	825.28	207.15	440.0 - 1245.00
Minimum fundamental frequency (Hz)	42	112.00	39.75	49.0 - 247.00	109.64	28.62	49.0 - 147.00
Fundamental frequency range (semitone)	42	34.89	10.76	3.0 - 56.00	35.07	6.88	24.0 - 50.00
Maximum SPL (dBA)	42	113.19	6.84	95.9 - 120.00	114.37	5.74	100.0 - 120.00
Minimum SPL (dBA)	42	77.10	10.39	58.0 - 95.10	78.60	11.57	50.8 - 98.7
SPL range (dBA)	42	36.76	10.50	16.6 - 59.20	44.93	42.54	18.9 - 255.00
Speaking Profile Area (dBA semitone)	42	423.69	197.36	36.5 - 873.50	451.89	215.72	93.3 - 1228.20

SPL= sound pressure level. dBA=sound pressure level measured in a-weighting scale. Hz=hertz

Voice Activity and Participation Profile (VAPP)

Since the VAPP for several subjects were incomplete, analysis was based on 41 subjects in the treatment group and 38 subjects in the no-treatment group.

Treatment Effect

Tables 9 and 10 list the descriptive statistics for the different VAPP sections for the treatment and no treatment groups respectively. For the treatment group, significant improvement was found for all the sections indicated by a decrease in scores across time, except for the Job section (see Appendix IV I). No significant changes were found for the no-treatment group (see Appendix IV I). To determine the extent of changes across time for the two subject groups, the difference in VAPP scores (Pre vs. Post-1 for the treatment group and Ax-1 vs. Ax-2 for the no-treatment group) were analysed. Significant differences were demonstrated for Self-perceived severity, Daily Communication ALS and PRS, Social Communication, total ALS, total PRS and total VAPP score. No significant difference was found for Job ALS and PRS and Emotion (see Appendix IV J).

Table 9. Descriptive statistics for the VAPP for the treatment group.

VAPP		Pre			Post-1			Post-2		
sections	**#N**	**Mean**	**SD**	**Range**	**Mean**	**SD**	**Range**	**Mean**	**SD**	**Range**
Self perceive severity	41	6.60	1.89	2.20-.60	3.68	2.01	0.60–8.30	3.38	2.42	0.20-0.00
Job	41	15.21	11.66	0.00-5.30	12.81	9.87	0.00-5.40	9.57	8.96	0.10-6.00
Daily	41	62.17	27.37	8.20-03.70	47.11	22.51	6.40-1.30	38.98	25.36	3.80-9.80
Social	41	14.04	11.79	0.00-5.70	9.08	8.68	0.00-2.50	8.36	8.08	0.00-4.80
Emotion	41	32.66	17.84	3.20-9.00	24.02	17.59	2.00-4.90	21.71	18.48	0.00-5.90
Total ALS	41	47.38	20.83	5.70-5.20	35.20	17.77	6.70-7.40	29.42	21.33	0.10-6.20
Total PRS	41	42.66	22.85	5.20-8.90	33.80	19.11	5.50-9.90	27.49	19.07	1.50-4.40
Total VAPP	41	130.92	58.20	20.90-45.30	97.50	50.68	16.60-0.60	82.00	27.28	2.20-39.60

*Significant at 0.01 level (two-tailed). #As one subject did not complete the VAPP, analysis was based on 41 subjects. Pre=pretreatment. Post-1=-posttreatment one. Post-2=posttreatment two. ALS= Activity Limitation Scores. PRS= Participation Restriction Score. Job=total ALS and PRS for Job. Daily=total ALS and PRS for Daily Communication. Social=total ALS and PRS for Social Communication.

Table 10. Descriptive statistics for the VAPP for the no-treatment group.

		Assessment 1			Assessment 2		
VAPP sections	**#N**	**Mean**	**SD**	**Range**	**Mean**	**SD**	**Range**
Self-perceived severity	38	6.17	2.42	0.90–10.00	5.74	2.31	1.10–9.50
Job	38	15.97	10.03	0.00–37.40	18.47	10.59	0.00–36.60
Daily	38	58.80	29.41	0.00–115.00	61.23	30.77	0.00–112.80
Social	38	15.97	12.48	0.00–47.30	16.34	10.89	0.00–37.90
Emotion	38	37.71	19.78	9.20–78.70	36.90	36.90	0.00–65.40
Total ALS	38	48.20	23.46	0.00–94.80	49.14	24.26	0.00–95.40
Total PRS	38	44.43	24.15	0.00–24.20	47.10	25.84	0.00–91.20
Total VAPP	38	135.70	165.33	1.80–256.70	137.88	69.86	1.10–260.70

*Significant at 0.01 level (two-tailed). #As 4 subject did not complete the VAPP, analysis was based on 38 subjects. ALS= Activity Limitation Scores. PRS= Participation Restriction Score. Job=total ALS and PRS for Job. Daily=total ALS and PRS for Daily Communication. Social=total ALS and PRS for Social Communication.

Maintenance Effect

A significant improvement was, however, found in total ALS and PRS for the Job section which did not demonstrate any significant improvement from Pre to Post-1 in the treatment group. Further analyses revealed that long term maintenance of treatment effect in which scores for the Self-perceived severity, total PRS, Daily Communication and Social Communication sections remained the same five weeks after treatment. In addition, there was significant improvement (decrease of scores) for total ALS and total VAPP (see Appendix IV I, Post-1 vs. Post-2 column).

Perceptual Voice Evaluation

Reliability

Since Judge 3 demonstrated poor intra-rater reliability of below 50% for agreement within one unit for roughness and overall severity, analyses were based on perceptual ratings from Judge 1 and Judge 2 only (see Appendix V A). Both Judge 1 and Judge 2 demonstrated intra-rater reliability of agreement within one unit at 84% or above. Judge 1 and Judge 2 demonstrated inter-rater agreement within one unit of above 60% for the three different voice qualities: roughness, breathiness and overall severity (see Appendix V B).

Treatment effect

Significant difference was demonstrated in the vocal quality across time between the treatment and no-treatment group in all the three voice quality parameters (see Table 13).

Table 13. Descriptive and inferential statistics for the perceptual voice evaluation of the treatment and no-treatment groups.

		Treatment			No-treatment				
Voice Quality	**N**	**Mean**	**SD**	**Range**	**Mean**	**SD**	**Range**	***t***	***p***
Roughness	42	1.71	1.06	0.00 -4.50	-1.73	1.25	-5.00 - 2.50	13.602	0.001*
Breathiness	42	1.79	1.68	1.50 -5.00	-1.54	1.35	-4.50 - 1.50	<u>10.005</u>	<u>0.001*</u>
Overall severity	42	2.36	1.29	0.50 -4.50	-1.89	1.38	-4.50 - 2.00	14.647	0.001*

Note. Negative mean value denotes worsening of vocal quality severity.

*Significant at 0.01 level (2-tailed). Results were based on Judge 1 and Judge 2's ratings only.

<u>Underlined</u> values were adjusted t values and p level where equal variances were not assumed due to significant results from Levene's Test for Equality of Variances using SPSS 10.0.

Maintenance Effect

Long-term treatment effective was demonstrated in 32 subjects in the treatment group based on roughness, breathiness and overall severity. Ten subjects demonstrated worsening vocal quality five weeks after treatment (see Appendix V C).

Accuracy in Classifying Vocal Quality Improvement After Treatment

Direct entry method of discriminant function analyses were carried out to determine the accuracy of the different independent (predicting) variables in predicting vocal quality changes after treatment using the selected measurements as reported in the Method section. The selected predicting variables were among the 14 outcome measures that had shown significant changes across time, they include 1) jitter %, 2) shimmer %, 3) NHR, 4) Pearson r at autocorrelation peak, 5) F_0 range 6) maximum F_0, 7) self perceived severity 8) daily communication ALS & PRS, 9) social communication ALS & PRS, 10) emotion, 11) total ALS, 12) total PRS, 13) total VAPP and 14) mean flow rate for sentence production. As there were only two groups for analysis for roughness (no change and improved vocal quality) based on Pre vs. Post-1 perceptual voice evaluation results, only one discriminant function was identified using 1) Pearson r at autocorrelation peak, 2) Maximum fundamental frequency and 3) Total VAPP as predictor variables. Table 14 lists the eigenvalues, canoncical correlations and Wilks' Lambda for voice quality changes in roughness after treatment.

Table 14. Canonical discriminant functions using the using the predictor variables into the different levels of vocal change for roughness and breathiness.

Dependent variables	Eigen-values	Canonical Correlations	Wilks' Lambdas	Chi- squares	df	*p*
Roughness	0.28	0.47	0.78	7.21	3	0.07
Breathiness Function 1	1.58	0.79	0.35	24.54	8	0.002*
Breathiness Function 2	0.10	0.30	0.91	2.27	3	0.52

Note. The predictor variables for roughness are Pearson *r* at autocorrelation peak, maximum fundamental frequency and total VAPP. The predictor variables for breathiness are Pearson *r* at autocorrelation peak, mean flow rate of sentence production, NHR and total VAPP.

*Significant at 0.01 level (two-tailed).

The accuracy in predicting group membership using the direct entry method can be evaluated by the percentage of correct classification of subjects in their original group and the eigenvalue. Eigenvalue refers to the between-group variance divided by the within-group variance. Powerful functions should have higher eigenvalue, ranging from 0 to 1. Hedderson (1991) claimed an eigenvalue of 0.40 to have sufficient discriminating power between groups.

As there were three groups for analysis for breathiness (worsened, no change and improved vocal quality) based on Pre vs. Post-1 perceptual voice evaluation results, two discriminant functions were identified using Pearson *r* at autocorrelation peak, mean flow rate for sentence production and Total VAPP as predictor variables. The eigenvalues, canoncical correlations and Wilks' Lambda for analysis of breathiness are presented in Table 14. The first discriminant function was significant (p=0.002) in classifying the subjects among the three levels of vocal changes. Function 1 accounted for 94% of the variance while function 2 accounted for the remaining 6%. Function 1 showed the highest correlation with mean flow rate for sentence production (r=-0.59) and Function 2 showed highest correlation with Pearson *r* at autocorrelation peak (r=0.82).

The overall classification rate using the selected predictor variables in classifying vocal changes in roughness was 78.80% and breathiness was 82.60% (see Table 15). The subjects with unchanged roughness and worsened breathiness were classified with a higher accuracy.

Table 15. Accuracy rate in classifying subjects using the predictor variables into the different levels of vocal change using discriminant function analysis.

Changes in voice quality	Group	No. of cases in original classification	No. of cases correctly classified	% of cases correctly classified	Overall Accuracy Rate
Roughness	Improved	31	24	77.40%	78.80%
	No change	2	2	100.00%	
Breathiness	Improved	23	19	82.60%	82.60%
	No change	3	2	66.70%	
	Worsened	2	2	100.00%	

Note. The predictor variables for roughness are Pearson *r* at autocorrelation peak, maximum fundamental frequency and total VAPP. The predictor variables for breathiness are Pearson *r* at autocorrelation peak, mean flow rate for sentence production, NHR and total VAPP.

DISCUSSION

Results revealed that the voice therapy program was effective in treating female dysphonic patients. This suggested that in treating voice disorders in female patients, using a variety of techniques including vocal hygiene, relaxation and facilitation exercises could reduce severity of dysphonia after therapy. There were significant differences between the two treatment groups in the amount of changes for the VAPP, perturbation measures (jitter and shimmer percent), auditory-perceptual evaluations, and maximum fundamental frequency. For the treatment group, there were additional changes from Pre to Post-1 state for fundamental frequency range and NHR. The following section provided separate discussions for the results on each of the measurements adopted in the present study.

Aerodynamic Analysis

It was shown that intra- and inter-subject variabilities in aerodynamic measures were high (Goozee et al., 1998; Higgins et al., 1994). The insignificant changes found in the treatment group might be a result of high variability in the aerodynamic measures between and within the subjects as evident from the large standard deviation values. The results found, therefore, might not be reliable in documenting vocal quality changes after treatment.

Acoustic Analysis

The results found was consistent with other efficacy studies in which acoustic measures (jitter %, shimmer % and NHR) could document significant changes across time with voice therapy (Carding et al., 1999; Fex et al., 1994; Roy, Bless, Heisey, and Ford, 1997; Verdolini-Marstson et al., 1994). VAM and STD were able to discriminate normal from dysphonic patients accurately (Callen et al., 1999; Fex et al.). But it was found in the present study that they were unable to document vocal quality changes after treatment. This might be because, though, vocal quality of the treatment group improved, the changes were not large enough to reach statistical significance and the vocal quality had not resumed to normal state, although the severity of dysphonia were reduced after treatment.

Since Hillenbrand's data analyses were only effective in evaluating dysphonic voices that deviate from normal in the direction of breathiness (Hillenbrand & Houde, 1996), the insignificant results might indicate that reduction in severity of breathiness after treatment was minimal. As breathiness was not always a feature in all dysphonic patients, this breathiness measures would not be able to detect voice quality improvement in subjects whose improvement was in other qualities, other than breathiness.

An interesting finding was that although there were no significant changes from Pre to Post-1 measurements for RPK and mean flow rate for sentence production, the treatment group demonstrated improvement in these two measures from Post-1 to Post-2 measurements. This might suggest that improvement in voice quality was not large enough to result in changes during Post-1 assessment. But the improvement was large enough to reach statistical significance when Post-2 measurements were taken for these breathiness and aerodynamic measures.

Voice Range Profile (Phonetogram)

Among the frequency and intensity measures, only maximum fundamental frequency and fundamental frequency range showed significant improvement from Pre to Post-1 measurements for the treatment group. This was because speaking in a wider fundamental frequency range and speaking in high frequencies required a more flexible phonatory system, regulated by vocal fold mass, length and tension, than speaking in low frequencies. Heylen, Wuyts, Merten, De-Bodt, Pattyn, Croux, and van de Heyning (1998) argued that some people with normal voices and dysphonic patients did not differ in their speaking lowest frequency values, so minimum fundamental frequency values might not be the most reliable indicator of voice quality improvement after treatment. On the other hand, the significant improvement in maximum fundamental frequency and fundamental frequency range should be resulted from reduction of laryngeal pathologies after therapy. With reduced vocal fold mass, the subjects were able to increase their speaking fundamental frequency and speak in a wider fundamental frequency range after treatment.

The finding that SPL measures were not decreased after therapy might suggest that the treatment was not quite successful to reduce loudness. However, it was possible that unchanged SPL measures was due to experimental situation and the subjects were required to do their recordings in a sound-proof room, which did not resemble everyday situation.

Voice Activity and Participation Profile (VAPP)

Results suggested that the subjects perceived improvement in their voice quality and reduction of impact of their voice problem on quality of life only after treatment was given to them. The lack of significant difference in the treatment group for the Job section was probably due to the fact that few of subjects had occupations that required high demand on voice usage. For example, among the subjects in the treatment group, five participants were teachers, four were sales, and one was a receptionist. Since the demand of voice usage for job was low for most subjects, it was expected that ALS and PRS for the Job section would be at the lower end of the scale and the difference in scores for pretreatment and posttreatment would remain relatively small.

The insignificant changes and significant improvement for Daily ALS and PRS and Total VAPP from Post-1 to Post-2 evaluations showed that subjects in the treatment group perceived that the impact of their voice problem on quality of life remained low during the maintenance phase five weeks after treatment.

Auditory-perceptual Evaluations

Findings from the perceptual evaluation indicated that roughness, breathiness and overall severity of the treatment group improved after voice therapy. On the other hand, mean values from the no-treatment group demonstrated deteriorated voice quality from Ax-1 to Ax-2 measurements (see Table 13 above).

The inter-rater reliability ranged from 63.50% to 69.00% for Judge 1 and Judge 2, this percentage of agreement was reported as acceptable as reviewed in the introduction. Since the listeners participated in the study were Year Four students from the Division of Speech and Hearing Sciences who underwent a 30-minute training program before the evaluation session, the level of inter-rater agreement was considered acceptable. Moreover, both Judge 1 and Judge 2 demonstrated good intra-rater reliability of 84% to 92% of agreement within one unit for the three types of voice quality. Although the training program did not include the quality of overall severity, the inter-rater reliability (69%) and intra-rater reliability (88%) suggested that overall severity was a reliable feature for perceptual voice judgement in documenting changes across time for the treatment and no-treatment groups.

An evaluation of all the measurements adopted suggested that the voice therapy program was effective in reducing severity of voice disorders for female dyphonic patients. Since no single measurement tool had been proved to be able to accurately diagnose voice disorders, it would be sensible to evaluate treatment efficacy from a variety of perspectives, including instrumental, clinicians's and patients' point of view.

Accuracy in Predicting Voice Quality Changes After Treatment

The second research question considered was to determine certain parameters that could predict voice quality improvement in terms of roughness and breathiness after treatment. Discriminant function analyses showed that the selected factors in predicting vocal changes in roughness and breathiness reached 78.80% and 82.10% respectively. The present findings for the accuracy in correct classification for roughness (78.80%) was quite high. However, careful examination of the canonical discriminant functions showed that the set of predictor variables (Pearson r at autocorrelation peak, maximum fundamental frequency and total VAPP) did not have strong discriminating power in predicting changes in roughness after treatment because eigenvalue (0.277) and canonical correlations (0.466) were low while Wilks' Lambda (0.783) was high. A powerful discriminating function should have high overall classification rate, eigenvalue and canonical correlations but low Wilks' Lambda. Moreover, since there were only two levels for the dependent variable (no change and improvement in roughness), an overall accuracy rate of 78.80% in classification of group membership was not particularly high. On the other hand, the selected variables (Pearson r at autocorrelation peak, mean flow rate for sentence production, NHR and total VAPP) for predicting changes in breathiness showed high discriminative power in predicting changes in breathiness after treatment. The large values of the eigenvalue (1.580), canonical correlations

(0.783) and the small value of Wilks' Lambda (0.352) further proved that the selected discriminating function was powerful.

In summary, results of the present study indicated that the voice therapy program was effective in treating voice problems for female dysphonic patients using a multi-dimensional approach of combining a number of voice measures in documentation voice quality improvement after treatment. The selected independent variables showed strong discriminating power in predicting changes in breathiness after treatment but not for roughness, this finding was probably due to the fact that changes in breathiness (as perceived from auditory-perceptual evaluations) correlated more with the pretreatment scores of the different outcome measures.

There were, however, limitations in the study. It was not unusual that subjects with unanalyzable voice samples using instrumental analysis were discarded from the subject group in studies using objective measures to assess voice disorders. However, unanalyzable data usually resulted from subjects with more severe voice disorders, the exclusion of these subjects would mean that subjects with severe voice disorder were excluded from analysis and valuable information would be lost from the study. Therefore, these unanalyzable data were preserved, although the variation in group size for several measurements, for example aerodynamic and acoustic analyses, would have affected the reliablity of the results found from the present study. For future larger scale studies, it was recommended to use the procedure described by Titze (1995) in categorizing the spectrogram of dysphonic subjects' voice samples into three different types. Based on this procedure, only Type I signals were included for acoustic analysis.

Another limitation relating to methodology was the amount of training given to the listeners for auditory-perceptual evaluations. A 30-minute training session proved to be not effective for certain listeners. One of listeners, Judge 3, produced poor intra- and inter-rater reliability, so perceptual ratings from Judge 3 had to be discarded from the final analysis. Seeing that the listening material for the actual auditory-perceptual evaluation was large, which required approximately two hours to rate all the voice samples, listeners were only trained for 30 minutes prior to the perceptual voice evaluation task. Future studies should provide more intensive training to student listeners or experienced speech therapist should act as listeners to increase the level of inter- and intra-rater reliabilities.

Clinical Implications

The present study revealed that the behaviorally based voice therapy program was effective in treating female dysphonic patients. In addition, the improvement in voice quality after treatment was demonstrated for 32 of the subjects in the treatment groups as revealed from perceptual voice evaluations. Since etiologies and types of dysphonia varied greatly among patients, in adopting the present voice therapy program to treat voice disorders, it was recommended that speech therapist should monitor each patient's needs and problems individually and render the treatment program flexibly for each patient.

As individual variation was large for measures like acoustic and aerodynamic analysis, it was recommended that assessment for voice problems and documentation of vocal quality changes following treatment in clinical situation should be based on a multidimensional approach using perceptual evaluation, functional measure, and voice range profile. Other instrumental measures should be used as an estimate or supplement to these measures and should not be used indiscriminately in clinical situation.

REFERENCE

Bielamowicz, S., Kreiman, J., Gerratt, B. R., Daucer, M. S., & Berke, G. S. (1996). Comparison of voice analysis systems for perturbation measurements. *Journal of Speech and Hearing Research, 39,* 126-134.

Callan, D. E., Kent, R. D., Roy, N., & Tasko, S. M. (1999). Self-organizing map for the classification of normal and disordered female voices. *Journal of Speech, Language, and Hearing Research, 42,* 355-366.

Carding, P. N., & Horsley, I. A. (1992). An evaluation study of voice therapy in non-organic dysphonia. *European Journal of Disorders of Communication, 27,* 137-158.

Carding, P. N., Horsley, I. A., & Docherty, G. J. (1999). A study of the effectiveness of voice therapy in the treatment of 45 patients with nonorganic dysphonia. *Journal of Voice, 13*(1), 72-104.

Fex, B., Fex, S., Shiromoto, O., and Hirano, M. (1994). Acoustic analysis of functional dysphonia: Before and after voice therapy (accent method). *Journal of Voice, 8*(2), 163-167.

Goozee, J. V., Murdoch, B. E., Theodoros, D. G., & Thompson, E. C. (1998). The effects of age and gender on laryngeal aerodynamics. *International Journal of Language and Communication Disorders, 33*(2), 221-238.

Hedderson, J. (1991). *SPSS / PC+ Made Simple*. Belmont, CA: Wadsworth-Thomson Publishing.

Heylen, L., Wuyts, F. L., Merten, F., De-Bodt, M., Pattyn, J., Croux C., & van de Heyning, P. H. (1998). Evaluation of the vocal performance of children using a voice range profile index. *Journal of Speech, Language and Hearing Research, 41*(2), 232-238.

Higgins, M. B., Netsell, R., & Schulte, L. (1994). Aerodynamic and electroglottographic measures of normal voice production: intrasubject variability within and across sessions. *Journal of Speech and Hearing Research, 37*(1), 38-45.

Hillenbrand, J, & Houde, R. A. (1996). Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. *Journal of Speech and Hearing Research, 39,* 311-321.

Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech and Hearing Research, 37,* 769-778.

Holmberg, E. B., Hillman, R. E., Hammarberg, B., Sodersten, M., & Doyle, P. (2001). Efficacy of a behaviorally based voiced therapy protocol for vocal nodules. *Journal of Voice, 15*(3), 395-412.

Jacobson, B. H., Johnson, A., Grywalski, C., Silbergleit, A., Jacobson, G. and Benninger, M. S. (1997). The Voice Handicap Index (VHI): Development and validation. *American Journal of Speech-Language Pathology, 6*(3), 66-70.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36,* 21-40.

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research, 35,* 512-520.

Ma, E. M. P., & Yiu, E. M. L. (2001). Voice activity and participation profile: Assessing the impact of voice disorders on daily activities. *Journal of Speech, Language, and Hearing Research, 44,* 511-524.

Pannbacker, M. (1998). Voice treatment techniques: a review and recommendations for outcome studies. *American Journal of Speech-Language Pathology, 7*(3), 49-64.

Roy, N., Bless, D. M., Heisey, D., & Ford, C. N. (1997). Manual circumlaryngeal therapy for functional dysphonia: An evaluation of short- and long-term treatment outcomes. *Journal of Voice, 11*(3), 321-331.

Roy, N., Gray, S. D., Simon, M., Dove, H., Corbin-Lewis, K., & Stemple, J. C. (2001). An evaluation of the effects of two treatment approaches for teachers with voice disorders: A prospective randomized clinical trial. *Journal of Speech, Language, and Hearing Research, 44,* 286-296.

Titze, I. R. (1995). *Workshop on Acoustic Voice Analysis: Summary Statement.* Iowa City: National Center for Voice and Speech.

Verdolini-Marston, K., Burke, M. K., Lessac, A., Glaze, L., & Caldwell, E. (1995). Preliminary study of two methods of treatment for laryngeal nodules. *Journal of Voice, 9*(1), 74-85.

Wilson, D. K. (1987). *Voice Problems in Children.* (3rd Ed.). Baltimore: Williams & Wilkins.

Yiu, E. M. L. (1999). Limitations of perturbation measures in clinical acoustic voice analysis. *Journal of Speech, Language and Hearing, 4,* 155-166.

Yiu, E. M. L., & Ma, E. P. M. (2001). Volume 2: Voice therapy instructional manual. In E. M. L. Yiu (Ed.), *Clinical Voice Assessment and Therapy: A Hong Kong Perspective.* Hong Kong: Sik Sik Yuen Education Research Fund/INSTEP, The University of Hong Kong.

Yiu, E., Murdoch, B., Hird, K., & Lau, P. (2002). Perception of synthesized voice quality in connected speech by Cantonese speakers. *Journal of Acoustical Society of America, 112*(3), 1091-1101.

Yiu, E., Worrall, L., Longland, J., & Mitchell, C. (2000). Analysing vocal quality of connected speech using Kay's computerized speech lab: A preliminary finding. *Clinical Linguistics & Phonetics, 14*(4), 295-305.

ACKNOWLEDGEMENT

I would like to thank my supervisor, Dr. Edwin Yiu, for his guidance and advice throughout the study. Deepest gratitude is also given to Dr. Estella Ma, for her kindest assistance, and Mr. Simon Lei, for his patience and generous help given during the course of the project. Last, but not least, I would like to thank the listeners for sacrificing their time to perform the perceptual voice evaluations in the study.

APPENDIX I

Occupational background of the subjects.

	Number of subjects	
Occupation	**Treatment**	**No-treatment**
Housewife	17	11
Teacher	5	7
Clerk	4	4
Sales	4	4
Student	2	1
Cleaner	2	1
Hawker	2	0
Receptionist	1	1
Nurse	1	1
School bus driver	1	0
Welfare worker	1	0
Court interpreter	1	0
Construction site supervisor	1	0
Cashier	1	0
Transportation manager	1	0
Social worker	0	2
Child care worker	0	1
Officer	0	1
Merchandiser	0	1
Dispenser	0	1
Unemployed	0	4

APPENDIX II

Pearson product-moment correlations of the pretreatment scores from the #14 outcome measures with changes in roughness and breathiness after treatment.

	Dependent variables	
Outcome measures	**Roughness (*p* level)**	**Breathiness (*p* level)**
Jitter %	-0.08 (0.63)	-0.53 (0.75)
Shimmer %	-0.23 (0.16)	0.14 (0.39)
Noise-to-harmonic ratio	-0.24 (0.14)	-0.03 (0.87)
Pearson *r* at autocorrelation peak	-0.14 (0.43)	0.42 (0.01*)
Fundamental frequency range	-0.15 (0.35)	0.05 (0.78)
Maximum fundamental frequency	-0.41 (0.009*)	0.01 (0.94)
Self perceived severity	0.19 (0.24)	-0.01 (0.93)
Daily communication ALS & PRS	0.19 (0.23)	0.20 (0.21)
Social communication ALS & PRS	0.19 (0.23)	0.26 (0.10)
Emotion	0.07 (0.62)	0.34 (0.03)
Total ALS	0.22 (0.17)	0.18 (0.25)
Total PRS	0.19 (0.24)	0.25 (0.12)
Total VAPP	0.19 (0.23)	0.28 (0.08)
Mean flow rate for sentence production	0.13 (0.45)	-0.07 (0.70)

*Significant at 0.01 level (two-tailed). #The 14 outcome measures are those that have shown significant differences across time for the treatment group. ALS=Activity Limitation Score. PRS=Participation Restriction Score. VAPP=Voice Activity and Participation Profile.

APPENDIX III

Discriminant function analysis summary using stepwise entry method for changes in voice quality (roughness and breathiness) after treatment.

Variable step number	**Variable entered**	***F* value to enter**	***p***
Roughness			
1	Pearson *r* at autocorrelation peak	5.80	0.02
Breathiness			
1	Mean flow rate for sentence production	6.99	0.004
2	Noise-to-harmonic ration	7.31	0.0001
3	Self perceived severity	6.77	0.0001

Note. The discriminant function analysis summary lists the actual variables entering the discriminant function equation, in the order of their effect on discrimination and the *F*-test value and appropriate significance level (*p*) at each step.

APPENDIX IV

A. Inferential statistics for aerodynamic measures of the treatment and no-treatment groups.

	No-treatment			Treatment				
	Assessment 1 vs. 2			Pre vs.Post-1			Post-1 vs. 2	
Aerodynamic measures	**N**	***t***	***p***	**N**	***t***	***p***	***t***	***p***
Mean flow rate (l/s)								
Maximum sustained /a/	41	0.56	0.58	42	-0.47	0.64	1.91	0.06
Maximum sustained /i/	42	-0.15	0.88	42	0.63	0.54	2.02	0.50
Maximum sustained /i/	38	0.31	0.76	42	2.01	0.05	-1.82	0.08
Most comfortable /a/	41	0.86	0.40	42	0.09	0.93	1.89	0.07
Sentence	39	–0.99	0.33	30	0.47	0.64	3.72	0.001*
Subglottal air pressure (cmH_2O)								
Sentence	22	0.05	0.96	16	-1.08	0.30	-0.41	0.69
Vowel-consonant strings	20	–2.18	0.04	25	-2.14	0.04	-0.70	0.50

*Significant at 0.01 level (two-tailed). Pre=pretreatment. Post-1=posttreatment one. Post-2=posttreatment two. l/s=litre per second. cmH_2O=centimeters of water.

B. Inferential statistics for the differences in aerodynamic measures between the treatment and no-treatment groups across time.

Aerodynamic measures	**Mann-Whitney U**	**Z score**	**p value**
Mean flow rate (l/s)			
Maximum sustained /a/	794.00	-0.61	0.54
Maximum sustained /i/	879.50	-0.02	0.98
Maximum sustained /i/	648.00	-1.45	0.15
Most comfortable /a/	809.00	-0.63	0.64
Sentence	455.50	-1.57	0.12
Subglottal air pressure (cmH_2O)			
Sentence	143.00	-0.98	0.33
Vowel-consonant strings	217.00	-0.75	0.45

l/s=litre per second; cmH_2O=centimeters of water

C. Inferential statistics for the acoustic measures (MDVP) of the treatment and no-treatment groups.

Acoustic	No-treatment			Treatment				
	Assessment 1 vs. 2				Pre vs. Post-1		Post-1 vs. Post-2	
measures	N	*t*	*p*	N	*t*	*p*	*t*	*p*
VAM	34	0.52	0.96	30	2.13	0.04	1.69	0.10
STD	34	2.63	0.01	30	-1.11	0.27	2.04	0.05
Jitter%	34	0.64	0.53	30	3.02	0.005*	1.88	0.07
Shimmer%	34	-0.85	0.40	30	3.37	0.002*	1.35	0.19
NHR	34	0.32	0.75	30	2.64	0.01	1.50	0.14

*Significant at 0.01 level (2-tailed). VAM=Coefficient of Variation of Peak to Peak Amplitude. STD=Standard Deviation of the Fundamental Frequency. NHR=Noise to Harmonic Ratio. Pre=pretreatment. Post-1=posttreatment one. Post-2=posttreatment two.

D. Inferential statistics for the differences in acoustic measures (MDVP) between the treatment and no-treatment groups across time.

Acoustic measures	Mann-Whitney U	Z score	p value
VAM	337.00	-2.33	0.02
STD	337.00	-2.33	0.02
Jitter %	423.00	-1.17	0.24
Shimmer%	317.00	-2.60	0.009*
NHR	312.50	-2.66	0.008*

*Significant at 0.01 (two-tailed). VAM=Coefficient of Variation of Peak to Peak Amplitude. STD=Standard Deviation of the Fundamental Frequency. NHR=Noise to Harmonic Ratio.

E. Inferential statistics for acoustic measures (Hillenbrand & Houde, 1996) for the treatment and no-treatment groups.

Acoustic	No-treatment			Treatment				
	Assessment 1 vs. 2				Pre vs. Post-1		Post-1 vs. Post-2	
measures	N	t	p	N	t	p	t	p
BRI	40	0.29	0.77	33	1.02	0.31	-1.82	0.08
H1H2	36	-2.36	0.02	38	-1.12	0.27	0.36	0.72
CPP	40	-0.14	0.89	35	-0.71	0.94	1.51	0.14
CPPS	40	-0.42	0.68	34	-0.56	0.96	2.63	0.01
RPK	39	-0.24	0.82	33	-0.15	0.88	3.05	0.004*

*Significant at 0.01 level (two-tailed). Pre=pretreatment. Post-1=posttreatment one. Post-2=posttreatment two. BRI=Breathiness Index. H1H2=First and Second Harmonics. CPP=Cepstral Peak Prominence. CPPS=Cepstral Peak Prominence-Smoothed. RPK=Pearson *r* at Autocorrelation Peak.

F. Inferential statistics for the differences in acoustic measures (Hillenbrand & Houde, 1996) between the treatment and no-treatment groups across time.

Acoustic measures	Mann-Whitney *U*	Z score	*p* value
BRI	654.00	-0.28	0.95
H1H2	425.00	-0.85	0.40
CPP	650.00	-0.73	0.60
CPPS	631.50	-0.61	0.60
RPK	532.50	-1.03	0.21

BRI=Breathiness Index. H1H2=First and Second Harmonics. CPP=Cepstral Peak Prominence. CPPS=Cepstral Peak Prominence-Smoothed. RPK=Pearson *r* at Autocorrelation Peak.

G. Inferenetial statistics for Voice Range Profile for the treatment and no-treatment groups.

Frequency and intensity measures	No-treatment: Assessment 1 vs. 2 N	*t*	*p*	Treatment N	Pre vs. Post-1 *t*	*p*	Post-1 vs. Post-2 *t*	*p*
Maximum F_0 (Hz)	42	0.47	0.64	42	-4.08	0.0001*	0.62	0.54
Minimum F_0 (Hz)	42	0.30	0.77	42	0.77	0.45	0.73	0.47
F_0 range (in semitone)	42	-0.10	0.92	42	-3.56	0.001*	-0.20	0.84
Maximum SPL (dBA)	42	-0.77	0.44	42	-2.30	0.03	-1.14	0.26
Minimum SPL (dBA)	42	-0.53	0.60	42	-2.53	0.02	-0.55	0.59
SPL range (dBA)	42	-1.04	0.31	42	0.29	0.78	-0.19	0.85
Speaking profile area (dBA semitone)	42	-0.72	0.48	42	-1.47	0.15	-0.07	0.52

*Significant at 0.01 level (two-tailed). Pre=pretreatment. Post-1=posttreatment one. Post-2=posttreatment two. F_0=fundamental frequency. SPL= sound pressure level. dBA=sound pressure level measured in a-weighting scale. Hz=hertz.

H. Inferential statistics for the differences in Voice Range Profile between the treatment and no-treatment groups across time.

Frequency and intensity measures	Mann-Whitney *U*	Z score	*p* value
Maximum fundamental frequency (Hz)	313.00	-3.08	0.002*
Minimum fundamental frequency (Hz)	506.50	-0.67	0.505
Fundamental frequency range (in semitone)	383.00	-2.21	0.027
Maximum SPL (dBA)	557.50	-0.31	0.975
Minmum SPL (dBA)	481.50	-0.98	0.33
SPL range (dBA)	556.50	-0.04	0.97
Speaking profile area (dBA semitone)	523.00	-0.46	0.65

*Significant at 0.01 level (two-tailed). SPL=sound pressure level. dBA=sound pressure level measured in a-weighting scale. Hz=hertz.

I. Inferential statistics for the VAPP of the treatment and no-treatment groups.

	No Treatment			Treatment				
	Assessment 1 vs. 2			Pre vs. Post-1			Post-1 vs. Post-2	
VAPP sections	**N**	***t***	***p***	**N**	***t***	***p***	***t***	***p***
Self perceived severity	38	1.27	0.21	41	7.97	0.0001*	0.82	0.42
Job	38	0.62	0.54	41	1.50	0.14	2.80	0.008*
Daily	38	-0.80	0.43	41	4.34	0.0001*	2.62	0.01
Social	38	-0.26	0.80	41	3.39	0.002*	0.79	0.44
Emotion	38	0.37	0.71	41	3.11	0.003*	2.12	0.04
Total ALS	38	-0.41	0.68	41	4.82	0.0001*	2.74	0.009*
Total PRS	38	-1.25	0.22	41	3.10	0.004*	2.44	0.02
Total VAPP	38	-0.38	0.70	41	4.41	0.0001*	2.87	0.007*

*Significant at 0.01 level (two-tailed). Pre=pretreatment. Post-1=-posttreatment one. Post-2=posttreatment two. ALS= Activity Limitation Scores. PRS= Participation Restriction Score. Job=total ALS and PRS for Job. Daily=total ALS and PRS for Daily Communication. Social=total ALS and PRS for Social Communication.

J. Inferential statistics for the differences in the VAPP between the treatment and no-treatment groups across time.

VAPP sections	Independent *t*	*p* value
Self perceived severity	5.00	0.0001*
Job ALS & PRS	1.55	0.126
Daily communication ALS & PRS	3.77	0.0001*
Social communication ALS & PRS	2.67	0.009*
Emotion	2.09	0.04
Total ALS	4.01	0.0001*
Total PRS	3.23	0.002*
Total VAPP	3.76	0.0001*

*Significant at 0.01 level (two-tailed). ALS=Activity Limitation Scores. PRS=Participation Restriction Score.

APPENDIX V

A. Intra-rater reliability for the perceptual voice evaluation.

	Voice Quality	Exact agreement	Agreement +/- 1 unit
Judge 1	Roughness	28.00% (7/25)	84.00% (21/25)
	Breathiness	40.00% (10/25)	88.00% (22/25)
	Overall severity	44.00% (11/25)	88.00% (22/25)
Judge 2	Roughness	60.00% (15/25)	76.00% (19/25)
	Breathiness	32.00% (8/25)	92.00% (23/25)
	Overall severity	60.00% (15/25)	88.00% (22/25)
Judge 3	Roughness	12.00% (6/25)	40.00% (10/25)
	Breathiness	36.00% (9/25)	64.00% (16/25)
	Overall severity	24.00% (3/25)	48.00% (11/25)

B. Inter-rater reliability for the perceptual voice evaluation.

Voice quality		Exact agreement	Agreement +/- 1 unit
Roughness	Judge 1 – Judge 2	30.20% (38/126)	68.30% (86/126)
Breathiness	Judge 1 – Judge 2	23.80% (30/126)	63.50% (80/126)
Overall severity	Judge 1 – Judge 2	30.20% (38/126)	69.00% (87/126)

C. Conditions of voice quality five weeks after voice therapy for the treatment group.

			Conditions of voice quality		
Voice Quality	N		Improved	No change	Worsened
Roughness	42	No. of subjects (%)	28 (66.67%)	4 (9.52%)	10 (23.81%)
Breathiness	42	No. of subjects (%)	25 (59.52%)	7 (16.67%)	10 (23.81%)
Overall severity	42	No. of subjects (%)	29 (69.05%)	3 (7.14%)	10 (23.81%)

Note. Conditions of voice quality were based on posttreatment one vs. posttreatment two perceptual voice evaluation results.